UNIVERSIDAD AUTONOMA DE MADRID

Escuela Politécnica Superior

19|20

Trabajo fin de grado

Técnicas de Machine Learning aplicadas a la predicción de los desvíos del Mercado Eléctrico

Rubén del Campo Hernando

Escuela Politécnica Superior
Universidad Autónoma de Madrid
C\Francisco Tomás y Valiente nº 11

excelencia Campus Internacional UAM CSIC

www.uam.es

UNIVERSIDAD AUTÓNOMA DE MADRID
ESCUELA POLITÉCNICA SUPERIOR

Grado en Ingeniería Informática

TRABAJO FIN DE GRADO

Técnicas de Machine Learning aplicadas a la predicción de los desvíos del Mercado Eléctrico

Autor: Rubén del Campo Hernando
Tutor: Álvaro Romero Miralles
Ponente: José Ramón Dorronsoro Ibero

mayo 2020

Rubén del Campo Hernando
Técnicas de Machine Learning aplicadas a la predicción de los desvíos del Mercado Eléctrico

Rubén del Campo Hernando
C\ Francisco Tomás y Valiente Nº 11

IMPRESO EN ESPAÑA – PRINTED IN SPAIN

A mi familia

“Someday we’ll look back on this and it will all seem funny.”.

Bruce Springsteen

Agradecimientos

En primer lugar, me gustaría agradecer al IIC la oportunidad de realizar este trabajo. Especialmente a Álvaro Romero que consiguió desatascar una situación que llevaba un rumbo bastante incierto. Nunca podré agradecertelo lo suficiente. Agradecer también a José Dorronsoro que se prestara tan de buen grado y pusiese tantas facilidades para ser mi ponente y a Julia Díaz por la confianza que depositó en mí desde el primer día.

Por supuesto, también se lo agradezco a mis padres y a mi hermano sin los cuales nunca hubiese llegado hasta aquí. Siempre me habéis apoyado en absolutamente todo, apostando siempre por la mejor formación para mí. Tampoco hay suficientes palabras de agradecimiento para vosotros.

Así mismo, me gustaría agradecer a mis compañeros del IIC lo mucho que me han enseñado en el tiempo que he pasado con ellos. No me olvido tampoco de todos mis compañeros de universidad con los que he disfrutado y sigo disfrutando de una gran compañía y amistad.

Resumen

El sistema eléctrico español, debido a su gran expansión, se ha visto obligado a establecer mecanismos de regulación y seguridad para el mercado eléctrico. En él encontramos una constante necesidad de equiparar la energía producida a la demanda. Por esta razón, las empresas generadoras deben dar cuenta de la energía que van a ser capaces de producir con anterioridad. El operador del sistema eléctrico (REE) se encarga de sancionar los casos en los que las productoras, yendo en contra del mercado, no se ajusten a sus predicciones. Este fenómeno es conocido como "desvío".

El objetivo de este trabajo es el de estudiar y comprender la gestión de estos desvíos y los precios de las sanciones que el operador impone, así como su predictibilidad. Para ello, se emplean modelos de clasificación para el caso de predicción del sentido desvíos y modelos de regresión para la predicción de sus precios. Además, tratando de establecerse como *baseline* para futuros trabajos a cerca de esta cuestión, se estudian y comparan dos formas de proporcionar los datos a los modelos: en *multi-output* (un modelo por horizonte de predicción) o en *single-output* (un solo modelo). Por lo tanto, este trabajo es en última estancia un *benchmark* de modelos con el fin de encontrar la predicciones más precisas.

Durante el desarrollo, se lleva a cabo el proceso de obtención, estudio estadístico y preparación de los datos con el objetivo de realizar, mediante técnicas de *Machine Learning*, las predicciones previamente explicadas y finalmente evaluar los resultados.

Para el primero de los problemas descritos, es decir, el sentido de los desvíos, se ha obtenido que la mejor solución consiste en emplear una *SVM* a la que se le presentan los datos en *multi-output*, obteniendo de este modo un 0.65 de *F1 score* y 0.73 de *precision score*.

En cuanto al problema del precio de los desvíos, se ha podido comprobar en la comparación que el resultado más preciso lo ofrece también la *SVM* con los datos presentados en *multi-output*, lo cual nos ofrece un 5.32€ de *MAE*.

Palabras clave

Sistema Eléctrico Español, Mercado Eléctrico Español, Operador del Sistema, Desvío, Predicciones, *Machine Learning*, Clasificación, Regresión

Abstract

The Spanish electricity system, due to its great expansion, has been forced to establish regulation and security mechanisms for the electricity market. In it we find a constant need to match the energy produced to demand. For this reason, generating companies must account for the energy that they will be able to produce previously. The electrical system operator (REE) is responsible for sanctioning cases in which the producers, going against the market, do not comply with their predictions. This phenomenon is known as "*deviation*".

The objective of this work is to study and understand the management of these deviations and the prices of the penalties that the operator imposes, as well as their predictability. For this, classification models are used in the case of prediction of the deviations and regression models for the prediction of their prices. In addition, trying to establish itself as a *baseline* for future work on this issue, two ways of providing the data to the models are studied and compared: *multi-output* (one model per prediction horizon) or *single-output* (a single model). Therefore, this work is ultimately a model *benchmark* in order to find the most accurate predictions.

The process of obtaining, statistical study and preparation of the data is carried out with the aim of making, through Machine Learning techniques, the previously explained predictions and finally evaluating the results.

For the first of the problems described, that is, the direction of the deviations, it has been found that the best solution is to use an *SVM* to which the data is presented in *multi-output*, thus obtaining a 0.65 of *F1 score* and 0.73 of *precision score*.

As for the problem of the price of the deviations, it has been possible to verify in the comparison that the most accurate result is also offered by the *SVM* with the data presented in *multi-output*, which offers us a 5.32€ *MAE*.

Keywords

Spanish Electricity System, Spanish Electricity Market, System Operator, Deviation, Classification,Regression, Machine Learning

Índice

Listas

Lista de figuras

Lista de tablas

1

Introducción

1.1. Motivación

El constante desarrollo de la sociedad mundial hacia un mundo en el que la tecnología está cada vez más presente en todos los ámbitos de la vida, ha traído consigo como consecuencia una transformación del sistema eléctrico hacia un modelo más competitivo, regulado y seguro. Concretamente en España este cambio fue visible a partir de enero de 1998, fecha en la que se llevaron a cabo los procesos legislativos necesarios para su liberalización. Desde ese momento, la complejidad del funcionamiento del sistema eléctrico español se incrementó notablemente, ya que al estar la generación de la electricidad en manos de empresas privadas que perseguirán siempre su propio beneficio, el Estado debe garantizar que estas empresas generadoras satisfagan en todo momento la demanda.

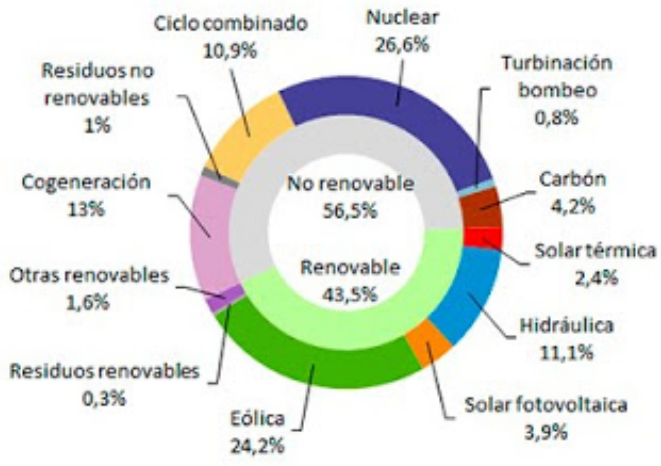


Figura 1.1: Desglose de las tecnologías de generación eléctrica en España en Marzo de 2019. Fuente: REE

Además, hay que tener en cuenta el gran auge de las energías renovables en el sistema eléctrico español en los últimos años. En la figura 1.1 podemos apreciar cómo la producción de la energía renovable supone ya un 40 % del total. Este tipo de energías, al estar sujetas a factores meteorológicos difícilmente previsibles como el viento o la radiación solar, resultan imposibles de planificar de forma

exacta.

Para llevar a cabo esta regulación del mercado eléctrico, se crearon diferentes mecanismos, empezando por la asignación del papel de único operador del sistema a Red Eléctrica de España (REE) [1], así como, los diferentes mercados de regulación primaria, secundaria, terciaria o la gestión de desvíos. Este trabajo pone el foco en este último punto.

La gestión de los desvíos se encarga de garantizar que la demanda y la producción de electricidad estén siempre equiparadas y, para ello, el operador del sistema convoca estos mercados con el fin de que las las empresas encargadas de la generación varíen su plan de producción de energía tratando de satisfacer las necesidades más inmediatas del sistema eléctrico. La necesidad de ajustarse a estos desvíos puede provocar grandes pérdidas económicas en las empresas generadoras de electricidad, ya que REE sanciona a aquellos que actúen en contra del sistema.

1.2. Objetivos

El objetivo global de este trabajo es el de estudiar y entender el fenómeno de los desvíos en el mercado eléctrico, tratando de predecir tanto su comportamiento como su precio a lo largo del tiempo, aplicando diversas técnicas de *Machine Learning*.

Para ello, el primer paso será comprender plenamente el mercado eléctrico español y la problemática que acarrea la gestión de la equiparación de la producción y la demanda. Esto nos permitirá asimilar qué son los desvíos, cómo se originan y cómo se comportan.

Una vez entendido el problema llevará a cabo el procedimiento clásico de un proyecto del ámbito de la Ciencia de Datos, basado en la metodología CRISP-DM [2]. Se empezará por la recolección de variables y el estudio estadístico de las mismas, para seguir con su tratamiento y procesado de los datos y para terminar con el uso de técnicas de *Machine Learning* en la predicción.

La pieza central de este trabajo será por lo tanto la parte referente al uso de algoritmos de *Machine Learning*. Se estudiarán los dos problemas más habituales en este campo: la clasificación y la regresión. La primera enfocada a predecir la ocurrencia de los desvíos y la segunda a predecir sus precios.

En ambos casos se realizará un *benchmark* de modelos donde se probarán diferentes modelos de distintas características (Random Forest, SVM, MLP, Logistic Regression) con el fin de compararlos. Asimismo, tanto en regresión como en clasificación se probarán dos formas distintas de proporcionar los datos a los modelos (*multi-output* y *single-output*) en la búsqueda de obtener así los mejores resultados para nuestro problema planteado.

1.3. Metodología CRISP-DM

Puesto que nos encontramos en un proyecto encuadrado dentro del ámbito de la Ciencia de Datos, se ha optado por emplear la metodología CRISP-DM para ordenar de forma secuencial las diferentes fases de las que constará. Por lo tanto, la estructura de las secciones del presente documento, seguirán de este modo dicha metodología, consistente en las fases enumeradas a continuación:

1.– Comprensión y entendimiento completo del problema planteado con el propósito de definir un objetivo concreto a lograr. Esta tarea será efectuada en profundidad en el capítulo 2 acerca del mercado eléctrico español.

2.– Elección de las variables que a priori podamos pensar que pueden influir en nuestro objetivo. El hecho de escoger bien estas variables se verá afectado en gran medida por nuestro grado de comprensión del problema alcanzado en la etapa anterior. En la las secciones 4.1 y 4.3 sobre las fuentes de datos y selección de variables se llevarán a cabo las explicaciones de este proceso.

3.– Estudio exploratorio de las variables recogidas mediante técnicas estadísticas utilizando como apoyo gráficas para poder entender visualmente las conclusiones obtenidas. Esto se realizará en la sección 4.2.

4.– Preprocesado de datos, construcción de nuestro *dataset* y definición del *pipeline* de predicción en las secciones 4.5 y 4.6, mediante las cuales conseguimos que las variables obtenidas puedan ser interpretadas correctamente por los modelos de predicción.

5.– Modelado a partir de las variables preprocesadas mediante el cual se emite una predicción para un momento determinado en el tiempo. Esto se explicará en la sección 5.

6.– Testing de las predicciones obtenidas y comparativa del desempeño de los distintos modelos en la sección 5.3.

1.4. Organización de la memoria

Este trabajo consta de seis capítulos:

- Capítulo 1: Motivación y objetivos del proyecto.
- Capítulo 2: El Mercado Eléctrico español donde se explica en profundidad la problemática de los desvíos.
- Capítulo 3: El estado del arte en relación a este tipo de problemas.
- Capítulo 4: Desarrollo central del proyecto donde se recogen, analizan y transforman los datos recolectados y se lleva a cabo la implementación de los modelos.
- Capítulo 5: Pruebas y resultados donde se explica el *benchmark* de modelos realizado y se compara la calidad de su desempeño.
- Capítulo 6: La conclusión final extraída del proyecto y las posibles vías de trabajo futuro.

2

Mercado eléctrico español

Tal y como se ha explicado previamente, en este proyecto empleamos la metodología CRISP-DM, cuyo primer paso consiste en entender el sector económico en donde se encuadra, es decir, el entendimiento del negocio. Deberemos conocer en profundidad todas las variables que pueden tener influencia, quién las proporciona, con qué frecuencia y cómo están calculadas.

En este caso, nos encontramos en un problema dentro del mercado eléctrico español. La electricidad es un componente fundamental en la sociedad, siendo impensable el modelo de vida actual sin la presencia de la misma en todos sus ámbitos. Es esta característica la que hace más interesante (y a su vez más complicado) nuestro problema, pues hace que el mercado eléctrico se vea afectado por una gran cantidad de factores independientes que pueden ser económicos, sociológicos, meteorológicos, etc. Por lo tanto, deberemos tener en cuenta todos estos factores para poder llegar a lograr un entendimiento completo del problema.

Durante toda esta sección se expondrá primero brevemente el funcionamiento general del sistema eléctrico español. Esa introducción servirá como base de conocimiento para poder así entender cómo funcionan sus distintos mercados en los que se compra y vende la energía. De este modo, podremos llegar a entender finalmente la problemática que intenta resolver este estudio en cuanto a la predicción de los desvíos.

Tomaremos como referencia principalmente las informaciones y conocimientos aportados por EYS [3], REE [1] y ESIOS [4].

2.1. Sistema eléctrico español

A pesar de que la historia el sistema eléctrico español se remonta desde mediados del siglo XX, no es hasta 1998 cuando, mediante la Ley 54/1997 [5] del Sector Eléctrico, se lleva a cabo un proceso de liberalización del sector y el sistema cobra una forma muy similar a la que tenemos hoy.

De este modo, todo el sistema eléctrico ha quedado separado en dos ámbitos claramente diferenciados: la gestión de la electricidad y la gestión económica.

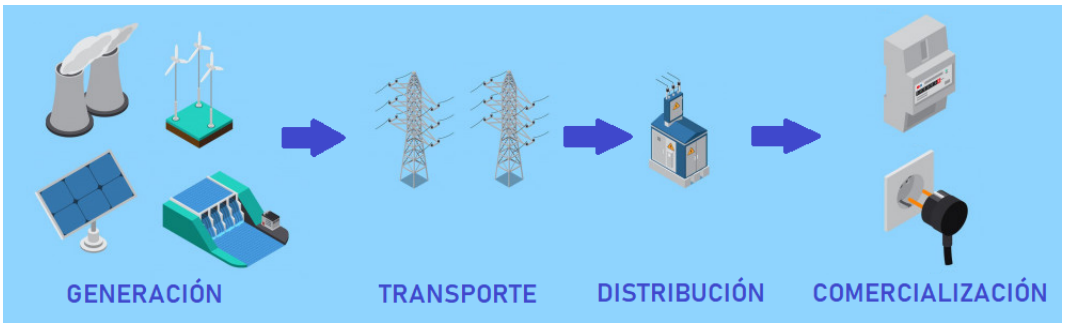


Figura 2.1: Fases de la electricidad en el Sistema Eléctrico. Fuente: Elaboración propia

En lo que se refiere a la gestión de la electricidad, encontramos las cuatro fases recogidas en la figura 2.1 que abarcan todo el recorrido desde que la energía es producida hasta que el consumidor final hace uso de ella. Estas cuatro fases son:

- **Generación**. Durante esta fase se lleva a cabo la producción de la energía eléctrica mediante las distintas tecnologías disponibles en el Sistema. Pueden abarcar desde las renovables como la eólica o la fotovoltaica, hasta no renovables como las que emplean combustibles fósiles o energía nuclear. Las empresas que se dedican a la generación, deben subastar esta energía en los distintos Mercados que se expondrán posteriormente.
- **Transporte**. Es la etapa durante la cual se transporta la electricidad a alta tensión desde las plantas de generación hasta las redes de distribución. El encargado de esta fase es REE.
- **Distribución**. Para esta fase, las empresas dedicadas a la distribución se encargan de hacer llegar a los consumidores la energía ya transformada proviniente de las líneas de transporte.
- **Comercialización**. Desde 1998, las labores de comercialización de la electricidad ya no depende de la compañía que suministra, pues esta es única para cada zona del territorio español. De este modo, a pesar de que sea una compañía concreta de distribución la que abastece a un consumidor, este puede acogerse a las distintas ofertas que le ofrecen las diversas empresas de comercialización. Por lo tanto, gracias al papel de las comercializadoras, se trata, a priori, de evitar la monopolización.

En cuanto a la parte de la gestión económica del sistema eléctrico, que en verdad es la que nos ocupa en este proyecto, será tratada en profundidad en los siguientes apartados del presente documento.

2.2. Mercado eléctrico español

En el mercado eléctrico español se lleva a cabo la compra y venta de energía eléctrica que es consumida o producida en territorio peninsular. Este gran mercado de flujo continuo debe ser dividido en dos partes: el mercado mayorista y el minorista.

El mercado mayorista se compone de una gran variedad de mercados donde sus agentes (productores, distribuidores, comercializadores y determinados consumidores) compran y venden energía eléctrica con una previsión que abarca desde el largo plazo de meses de antelación, hasta el corto

plazo de apenas horas. Todos estos mercados están agrupados dentro del **MIBEL** [6] (Mercado Ibérico de la Electricidad), el cual también gestiona los mercados portugueses. Dentro de los mercados del MIBEL caben destacar el **OMIE** [7] (Operador del Mercado Ibérico Española), del cual dependen los mercados destinados al corto plazo, y el OMIP (Operador del Mercado Ibérico de Energía - Polo Portugués), del que dependen los mercados de largo plazo.

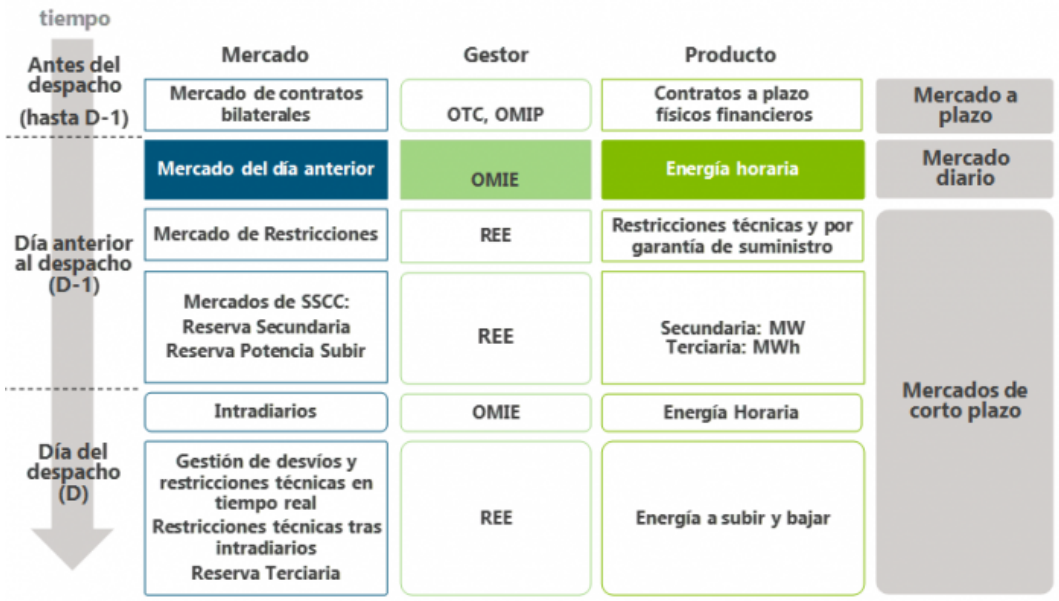


Figura 2.2: Secuencia de mercados en el Mercado Ibérico de Electricidad (MIBEL). Fuente: Energía y Sociedad [3]

En la figura 2.2, podemos observar a modo de resumen todos los distintos mercados que se expondrán posteriormente, así como el momento en el tiempo en el que acontecen y el Operador que los gestiona.

El principal mercado gestionado por el OMIE es el llamado "**mercado diario**" o "spot diario", en el cual los agentes presentan sus ofertas de venta y de compra antes de las 14:00 horas del día D, para todas las horas que componen el día D+1. Las ofertas de compra las efectúan los agentes de comercialización y las de venta los de generación. El sistema utilizado es llamado "marginalista", en el cual los ofertantes ponen a la venta una cantidad de energía a un determinado precio menor o igual al llamado **"precio de casación"**, para la cual otros agentes realizan ofertas de compra a un precio mayor o igual al "precio de casación". Este se calcula como la intersección entre las curvas de oferta y demanda para una hora en concreto.

Las cantidades de energía que se subastan están sometidas a las restricciones físicas de cada planta, así como de la capacidad de la red de transporte. Cabe destacar en relación a esto que según las reglas del mercado, las empresas generadoras, están obligadas a ofertar toda la capacidad disponible de sus instalaciones a lo largo de toda la secuencia de mercados.

El precio al que los agentes ofertan incluye desde el precio al que evitarían incurrir si decidiesen

no producir (costes de arranque, mantenimientos, etc.), así como los costes derivados por el hecho de producir (coste de oportunidad).

A partir del momento en que los distintos agentes han puesto sus ofertas en el mercado para cada una de las horas del próximo día, el operador OMIE se encarga de ordenarlas de menor a mayor precio. De este modo, las energías eólicas y fotovoltaica entran casi siempre a precio nulo, seguidos por la nuclear, las hidráulicas fluyentes y los ciclos combinados, para encontrar por último a las hidráulicas regulables [8].

Además del mercado diario, el OMIE también se encarga de gestionar los diferentes **mercados intradiarios**. Estos mercados se caracterizan por producirse durante el mismo día en que se despacha la energía, aunque su operativa es similar a la del mercado diario. En este caso, se establecen seis sesiones a lo largo del día, que los ofertantes aprovechan para reajustar sus posiciones tomadas previamente a partir de la última información que van recibiendo, lo que les permite ser más precisos en sus estimaciones o incluso llegar a sobrellevar incidencias técnicas en el sistema. Las seis sesiones las encontramos a las 17:00, 21:00, 01:00, 04:00, 08:00 y 12:00. Estas ofertas duran unos cuarenta y cinco minutos, a excepción del primer intradiario que es de ciento cinco minutos.

Todos los resultados obtenidos de las diversas transacciones deben de ser conocidos y verificados por el Operador del Sistema (REE) después de que el OMIE resuelva la casación igualando la oferta a la demanda prevista. Antes de llevar a cabo la aceptación de estas compras, REE deberá de asegurarse de que son factibles desde el punto de vista técnico de tal forma que se asegure la integridad y buen funcionamiento de toda la red de transporte y suministro. En el caso de que se detecte una de estas "*restricciones*", entran en juego los llamados "***mercados de ajuste***", mediante los cuales, REE trata de solventar la contingencia mediante el uso de ciertos grupos de generación, intentando siempre de incurrir en el menor coste. Esto implica que debe de disponerse siempre de grupos de generación de reserva que puedan iniciar inmediatamente la generación suplementaria en caso de que el sistema lo requiera. En estos mercados de ajuste, todas las compras son efectuadas por REE como operador del sistema, que puede comprar a cualquier generador que desee vender en este tipo de mercados.

Dentro de los mercados de ajuste podemos encontrar tres mercados operados por REE:

- **Gestión de restricciones técnicas**. Se encarga de la descongestión de la red de transporte y distribución. Este mercado tiene su actuación tanto en el día previo al despacho de la energía, como del propio día en curso tras el cierre del mercado intradiario. El equipo técnico de REE, gracias a sus herramientas de medición y simulación, es capaz de determinar el comportamiento de la red, por lo que hará uso de este mercado para solventar las congestiones detectadas modificando el programa de generación inicial.
- **Gestión de los servicios complementarios**. Tiene como cometido el de equiparar en todo momento la generación con la demanda. Para ello hace uso de los mercados de regulación primaria, regulación secundaria y regulación terciaria, ordenados de menor a mayor necesidad de tiempo de respuesta en la generación.
- **Gestión de desvíos**. Este mercado, al ser el que nos ocupa en este estudio, dispondrá de un apartado propio posteriormente, donde se estudiará más en profundidad su causa, cálculo, comportamiento y consecuencias.

2.2.1. Mercado de gestión de Desvíos

Como se ha explicado previamente, el Operador del Sistema (REE), tiene como cometido el de equilibrar escrupulosamente el consumo con la demanda [9]. A pesar de que esta tarea se empieza a perfilar con muchas horas de antelación antes despacho de la energía mediante los mercados de largo plazo, el mercado diario, los intradiarios y los mercados de ajuste mencionados en la sección anterior, se dispone de un último mecanismo que es la gestión de desvíos.

Figura 2.3: Tabla resumen de los posibles casos que pueden darse en la liquidación de la gestión de los desvíos. Fuente: Elaboración propia

Los distintos generadores deben comunicar constantemente a REE cuánto prevén que se van a desviar de sus predicciones de generación. Estas medidas, junto con una predicción interna que lleva a cabo REE de la generación de energías renovables y de demanda, pueden provocar que si se detecta un desvío superior a los 300 MW de media horaria, se decida convocar el mercado de gestión de desvíos. Mediante este mercado, el Operador del Sistema se encarga de pedir ofertas a los distintos generadores de forma que sean en el sentido contrario a los desvíos detectados en sus cálculos. Encontramos por lo tanto, de acuerdo a EYS [3], dos tipos posibles de desvíos:

- **Desvíos a subir**. Se producen cuando las generadoras originalmente sobreestiman lo que finalmente van a ser capaces de producir. En este caso se piden ofertas para incrementar la producción.
- **Desvíos a bajar**. Se producen cuando las generadoras originalmente subestiman lo que finalmente van a ser capaces de producir. En este caso se piden ofertas para reducir la producción.

En la figura 2.4 podemos ver cómo el sentido de los desvíos oscila a lo largo de las diferentes horas dentro de una semana. A través de esta información, conseguimos saber si en cada una de esas horas REE declaraba un déficit o un superávit de energía en el Sistema.

En lo que se refiere al punto de vista económico, debemos ahondar en la liquidación de estos envíos, pues es en el esta problemática en donde se sustenta la motivación final de este estudio.

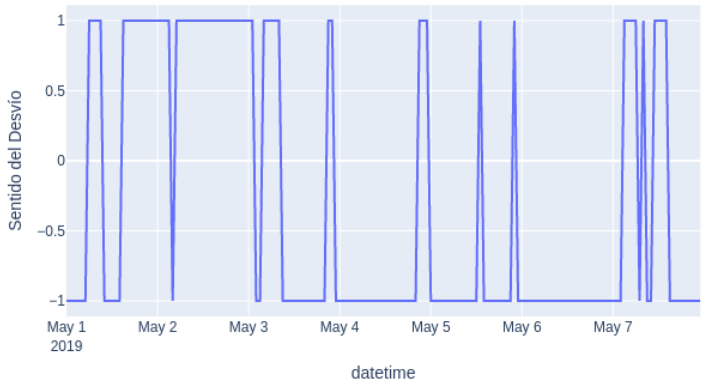


Figura 2.4: Evolución del sentido de los desvíos a lo largo de una semana. Fuente: Elaboración propia a partir de datos de ESIOS [4]

REE, en su papel de Operador del Sistema, tendrá como cometido el de sancionar a todos aquellos que, durante la celebración del mercado de gestión de desvíos, se hayan comportado en contra del mercado, es decir, que lo hayan perjudicado. Concretamente el porcentaje que se cobra de menos o se paga de más respecto al precio del mercado diario vendría dado como:

$$CosteDesvContrario = 100 * (PrecioDesvBajar - PrecioDesvSubir)/PrecioDiario$$

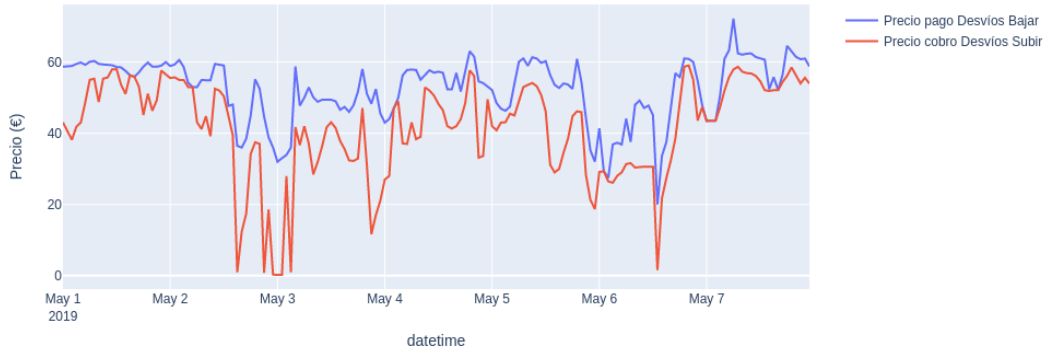


Figura 2.5: Evolución del precio de los desvíos a lo largo de una semana. Fuente: Elaboración propia a partir de datos de ESIOS [4]

En la figura 2.3 podemos observar los 4 posibles casos que pueden darse según la forma de producir energía por parte de los generadores en función de las necesidades del mercado:

1.– En caso de que se finalmente produzca más de lo previsto cuando el mercado requiere más energía programada, el generador cobrará a precio de mercado diario toda la energía extra que ha generado.

2.– Si se produce finalmente más de lo previsto cuando el mercado demanda menos energía de la programada, se penaliza al generador cobrando menos de lo que hubiera cobrado si hubiera venido esa energía en el mercado diario. Este precio se calcula como el mínimo entre el precio del mercado diario y el precio medio ponderado de las energías a bajar de RR, regulación terciaria y regulación secundaria.

3.– Cuando se produce finalmente menos de lo previsto y el mercado demanda más energía que la programada, el generador pagará a precio de desvío la energía que nunca ha llegado a generar. Este precio se calcula como el máximo entre el precio del mercado diario y el precio medio ponderado de las energías a subir de RR, regulación terciaria y regulación secundaria.

4.– En el caso en que finalmente se produce menos de lo previsto cuando el mercado requiere menos energía que la programada, no ocurre nada. Simplemente no cobrará nada por la energía que no ha llegado a generar.

En la figura 2.5 podemos observar la evolución del precio de los desvíos a subir y a bajar. Como es de esperar, ya que, como se ha enunciado previamente, el primero va a ser siempre menor o igual que el segundo, puesto que uno se calcula a partir del mínimo del precio diario y los precios de regulación, mientras que el otro se calcula aplicando el máximo. Se aprecia que los precios presentan una variación bastante alta en el tiempo debido a la gran multitud de factores que los influyen, por lo que su predicción no será sencilla.

3

Estado del arte

A lo largo de este capítulo se mostrará el marco teórico que en el que se encuadra el proyecto que nos ocupa, así como los distintos trabajos previos que nos servirán como referencia y punto de partida en el desarrollo.

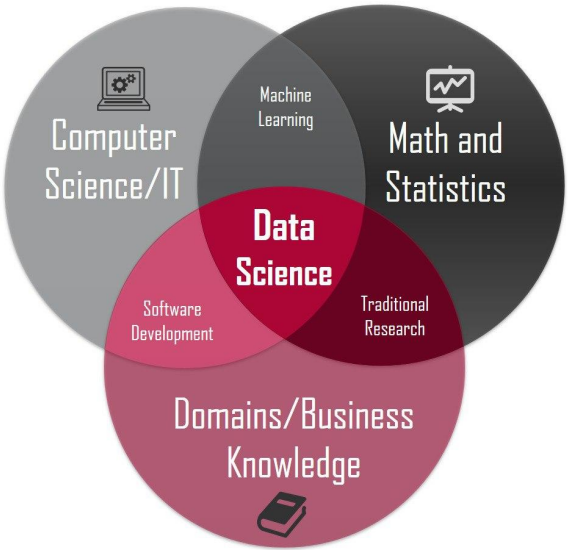


Figura 3.1: Esquema de ramas de conocimiento dentro de la Ciencia de Datos. Fuente: TowardsDataScience [10]

Este trabajo se encuentra dentro del campo de la Ciencia de Datos, el cual tiene la peculiaridad de estar sustentado sobre tres ramas de conocimiento que históricamente estaban separadas. Dicha característica se aprecia claramente en la figura 3.1. Por un lado hay una parte de comprensión del problema y del área de negocio en donde se encuadra, que en nuestro caso es el mercado eléctrico español [2]. Por otro lado, es necesaria una base matemática y estadística que nos ayude a comprender los datos recolectados y elaborar modelos de predicción que resuelvan nuestro problema. Por último, es indispensable disponer de conocimiento sobre Ciencias de la Computación, pues necesitamos programar algoritmos que recojan los datos y los procesen.

3.1. Trabajos previos

Tal y como se ha explicado extensamente durante el capítulo dedicado al mercado eléctrico español 2.2.1, la gestión de los desvíos es un problema bastante complejo. Esto, unido al hecho de que es un problema muy concreto del mercado español, ha supuesto que apenas ha podido ser posible encontrar trabajos anteriores que aborden este mismo problema.

En lo referente a la predicción del sentido de los desvíos se ha tratado en [11]. En este trabajo se ha estudiado la predictibilidad de los desvíos desde un punto teórico e hipotético, ya que finalmente no se llega a ninguna implementación de la solución. Se propone el uso un modelo bastante simple de red neuronal, que es el Perceptrón Multicapa (*MLP*) con una arquitectura de solamente una capa oculta de 25 neuronas. Como se ha mencionado, no llega a ninguna implementación final, por lo que no podemos utilizarlo como referencia en la comparación de resultados.

Debido a la no existencia de trabajos similares al que nos ocupa dentro del ámbito académico, se ha decidido ampliar el marco de búsqueda a otros problemas parecidos. Podemos llegar a afirmar de este modo, que el problema que se nos plantea es el del ***time-dependent classification***, ya que establecemos que para cada hora Y en que queremos predecir el sentido, tenemos una X formada por toda la serie temporal que abarca hasta dicha hora. Es decir, el sentido se ve influido por los momentos pasados.

Sobre este campo de estudio se han propuesto soluciones como se detallan en [12]. Propone a grandes rasgos utilizar para los distintos Hn horizontes, Cn clasificadores. Para predecir el horizonte H1 usará el clasificador C1; para el H2, un ensemble del C1 y C2; para el H3, un ensemble del C1, C2 y C3, y así se sigue el procedimiento hasta el último horizonte que se desee predecir. El ensemble tendrá implementado un sistema de votación para elegir la clase que más veces ha sido predicha por los distintos clasificadores. Gracias a este tipo de implementaciones se puede lograr que la predicción sea dependiente de los tiempos anteriores.

Como ya se ha visto en el capítulo 2.2.1 acerca de los desvíos en el mercado eléctrico, es importante fijar un horizonte de predicción. En nuestro caso, lo hemos establecido las 0:00 horas del día D, y se harán predicciones para las siguientes 24 horas, es decir, desde la 1:00 hasta las 00:00 del día D+1. Esta idea se explicará más en detalle en la próxima sección 4 *Desarrollo*, pero podemos resumirla en que, puesto que en cada ocasión queremos predecir 24 targets (las siguientes 24 horizontes cada día), se ha entendido como un problema de *time-dependent classification*, y más concretamente es ***multi-output classification***. La técnica consiste en utilizar un modelo de clasificación distinto para cada target, es decir, un modelo para cada hora. Este paradigma está basado en el trabajo desarrollado en [13].

Una vez definido lo que es un *multi-output classifier*, podemos entender que cada uno de esos modelos individuales puede ser cualquiera de nuestra elección. En nuestra búsqueda de la mejor pre-

dicción, llevaremos a cabo pruebas sobre distintos tipos de modelos de clasificación. La base teórica sobre la que se fundamentan los algoritmos que se emplearán, la podemos encontrar en [14].

Por otro lado, se harán también pruebas de estos mismos modelos en una versión sin *multi-output*, es decir, tendremos un un solo target por cada fila del dataset de train. A esta versión la llamaremos *single-output*.

En cuanto a lo que se refiere al problema de la predicción de los desvíos, es un problema de regresión al uso, que no encierra tanta complejidad como el de clasificación y podemos servirnos de más referencias.

A pesar de que la predicción de los precios de los desvíos tampoco es un problema que se haya podido encontrar como estudiado en trabajos previos, se puede entender que es un problema parecido al de la predicción del precio de la electricidad en el mercado diario. Este sí que es un problema bastante más estudiado y sobre el que podemos encontrar más publicaciones como en [15] o en [16]. El primero utiliza algoritmos clásicos de regresión como *Random Forest* o *SVM*, mientras que el segundo utiliza la *ARIMA* (*Autoregressive Integrated Moving Average*), que es un modelo dinámico de gran utilidad cuando hablamos de series temporales. Aunque ambos nos suponen un buen punto de referencia, nos basaremos en las ideas del primero para llevar a cabo el desarrollo de este trabajo.

3.2. Modelos de predicción

Uno de los modelos utilizados será la ***Logistic Regression*** (*LR*) [14], que a pesar de su nombre, solo tiene cabida dentro de los problemas de clasificación. Se fundamenta en el intento de correlacionar la probabilidad de un target binario con una variable escalar. Es decir, que tratará de aproximar la probabilidad de que el target sea 1 ó 0 dado el valor de la variable.

Otro de los modelos empleados será el ***Random Forest*** (*RF*) [14], basado en la técnica de bagging. A grandes rasgos, podemos afirmar que se genera una gran cantidad de árboles donde cada uno utiliza una cantidad reducida y aleatoria de variables respecto al total del conjunto de entrenamiento. Todos los árboles crecen hasta la misma longitud, quedándose con el mejor por votación en los problemas de clasificación y con un promedio en el caso de la regresión.

Utilizaremos también el modelo ***Support Vector Machine*** (*SVM*) [17], claramente el más complejo desde el punto de vista matemático. Intuitivamente podemos decir que la idea sobre la que se fundamenta es en la de proyectar nuestros datos a un espacio de una dimensionalidad superior en donde se define un hiperplano o conjunto de hiperplanos que tengan máxima distancia respecto a los puntos de las clases a predecir. También cumple las características para ser usado dentro de nuestro problema de regresión.

También emplearemos el ***Multi Layer Perceptron*** (*MLP*) [18], modelo basado en la arquitectura

de red neuronal, es decir, neuronas agrupadas en capas y conectadas entre sí. El hecho de disponer de múltiples capas, le permite resolver problemas no separables linealmente. La capa de entrada tiene tantas neuronas como variables contenga nuestro modelo; y la capa de salida, tantas neuronas como targets a predecir. El número de capas y neuronas intermedias (llamadas *capas ocultas*), así como las funciones de activación de dichas neuronas, deben ser configuradas del modo que mejor resultados ofrezca, sin olvidar del alto coste computacional que este suponga.

Todos estos modelos, como ya se ha explicado en la sección anterior, serán incluidos dentro de un *multi-output classifier* para que se haga un modelo distinto por hora del día a predecir. Sin embargo, algunos de ellos como el *Random Forest* o el *Multi Layer Perceptron*, también ofrecen la posibilidad de funcionar como *multi-output* sin la necesidad de incluirlos dentro de un *multi-output classifier*.

3.3. Métodos de evaluación de resultados

A pesar de que, como se ha recalcado previamente, no podemos comparar nuestros resultados con trabajos previos sobre el tema, es necesario que definamos una serie de métricas, de acuerdo con [19], que nos permitan evaluar la calidad de las predicciones efectuadas por los distintos modelos desarrollados.

En el caso de la predicción del sentido de los desvíos, al ser un problema de clasificación se han decidido utilizar tres métricas bastante habituales para este tipo de problemas: *precision*, *recall* y *F1*. Para entender qué evalúan estas métricas, es necesario comprender previamente los siguientes conceptos inherentes a la clasificación. Suponiendo que en un problema de clasificación binario (A, B) queremos predecir la clase A, tenemos los siguientes conceptos:

- ***True Positive*** (**TP**) cuando se ha predicho A para un elemento de clase A.
- ***False Positive*** (**FP**) cuando se ha predicho A para un elemento de clase B.
- ***False Negative*** (**FN**) cuando se ha predicho B para un elemento de clase A.
- ***True Negative*** (**TN**) cuando se ha predicho B para un elemento de clase B.

Teniendo en presentes estos conceptos, podemos pasar a enunciar las distintas métricas que vamos a emplear en clasificación:

- ***Precision*** consiste en efectuar el ratio de TP / (TP + FP). Intuitivamente podemos considerarla como la capacidad del clasificador de no etiquetar como A un elemento de clase B.
- ***Recall*** se calcula con el ratio TP / (TP + FN). Podemos decir que intuitivamente supone la capacidad del clasificador de etiquetar correctamente todos los elementos A.
- ***F1*** es la media harmónica de *precision* y *recall*. De este modo, podemos evaluar que nuestro modelo se comporte bien en ambas métricas simultáneamente.

Todas estas métricas mencionadas abarcan valores entre 0 y 1, donde el 1 es el mejor resultado

posible y 0 el peor.

Para el caso de los problemas de regresión, como en el que nos encontramos a la hora de predecir el precio de los desvíos, debemos utilizar otro tipo de métricas, ya que el propósito no es saber ya la calidad del modelo a la hora de etiquetar clases, sino el de saber lo cerca que se ha quedado de predecir una cantidad concreta. Las métricas que utilizaremos son:

- ***MAE*** (***Mean Absolute Error***) que se calcula como: $\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)$, donde los y_i son los valores predichos por nuestro modelo, los x_i los valores de nuestro conjunto de test con los que comparamos la calidad de nuestras predicciones y n el número de patrones que se quieren testear. Intuitivamente podemos decir que nos proporciona cuánto error tienen de media nuestras predicciones.
- ***MSE*** (***Mean Squared Error***), calculado como $\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2$. Se puede decir que nos indica cuánto error medio tienen nuestras predicciones acentuando más los errores más grandes.

4

Desarrollo

Una vez se ha entendido el concepto de desvío y la problemática que acarrean, llega el momento de intentar encontrar una solución que nos ayude a predecir el sentido en el que ocurren, así como el precio asociado a ellos.

A lo largo de esta sección se recorrerán las distintas partes indispensables que se han considerado como indispensables en el desarrollo de la resolución del problema. Este camino pasará por buscar las distintas variables que pueden servirnos para explicar el suceso, la forma de obtenerlas y los predictores implementados, la manera de construir el *dataset* de entrenamiento con ellas, así como el *pipeline* de predicción, para acabar con el módulo que nos ayude a testear la calidad de nuestras predicciones.

4.1. Fuentes de datos

Puesto que nos encontramos haciendo un estudio sobre un fenómeno que ocurre dentro del Mercado Eléctrico español, nuestra principal fuente de datos será el **ESIOS** [4] (Sistema de Información del Operador del Sistema). Es un sistema de información desarrollado por Red Eléctrica de España con el fin de poner de forma transparente al alcance de cualquier usuario una gran cantidad de información casi en tiempo real acerca de la generación de energía, la demanda o el estado de los distintos mercados. Además, dispone de una API REST pública mediante la cual el programador puede obtener mediante peticiones HTTPS obtener datos en formato JSON sobre una variable en un rango de tiempo deseado.

Para desarrollar la solución planteada se ha implementado un módulo de obtención de datos mediante el cual se ha podido trabajar cómodamente en las fases posteriores a la obtención de datos, ya que gracias a él es trivial añadir una nueva variable que puede ser considerada importante.

Por otro lado, después del estudio del mercado eléctrico realizado previamente, cabe pensar que los desvíos, al tener una relación estrecha con la demanda de energía, pueden verse influidos por los días festivos, ya que la demanda es muy distinta en este tipo de días. Es por este motivo por el que se ha decidido recopilar los días festivos en España, y para ello se ha hecho uso de la API

REST pública proporcionada por ***Nager.Date*** [20]. Como las fechas de las festividades por sí solas no nos aportan demasiada información, se han relacionado con la cantidad de personas que viven en la Provincia o Comunidad Autónoma en la que hay un festivo. Estos datos se han obtenido del **INE** [21] (Instituto Nacional de Estadística). De este modo, la variable final que hemos obtenido es el porcentaje de población española que se encuentra en un día festivo para cada día de un año. Métodos similares han sido aplicados en trabajos previos como [15].

4.2. Análisis exploratorio y estadístico

Tal y como hemos hablado en el capítulo sobre el mercado eléctrico español 2, nuestro objetivo será el de predecir el sentido de los desvíos.

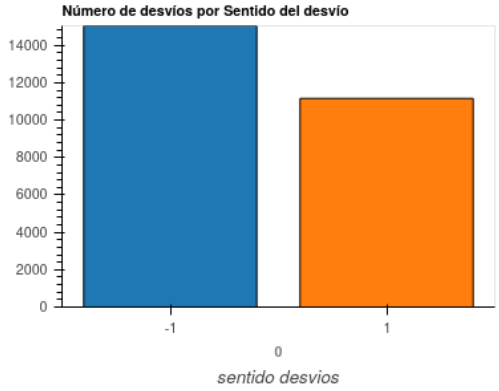
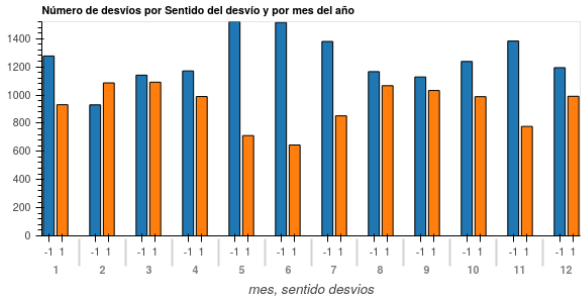


Figura 4.1: Distribución del sentido de los desvíos. Fuente: Elaboración propia a partir de datos de ESIOS

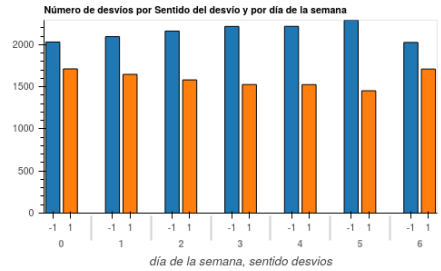
Primero contabilizaremos la cantidad de sentidos positivos y negativos que encontramos en nuestro conjunto de datos. Podemos observar en la figura 4.1 que el 57 % de las horas encontramos un sentido negativo, mientras que el el 43 % vemos un sentido positivo. A pesar de que nuestro target está desbalanceado, no es una diferencia significativa que nos obligue a emplear técnicas para reducirlo.

A continuación, seguimos estudiando cómo se comportan a lo largo de los tres años de datos horarios (en UTC) que utilizaremos en este estudio aplicando agrupaciones de datos según diferentes intervalos temporales.

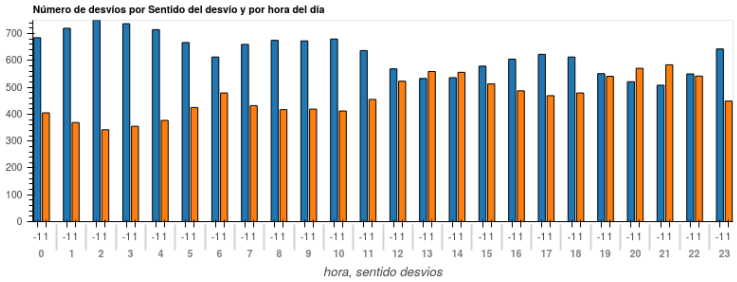
En todas las figuras de 4.2 podemos observar que el número de desvíos por sentido tiene una variabilidad bastante alta a lo largo del tiempo. Si fijamos la atención en la gráfica 4.2(c) que agrupa por hora del día, vemos que para la primera mitad del día hay una diferencia bastante amplia entre la ocurrencia de sentidos positivos y negativos, pero una vez pasamos de las 12 del medio día, tiende a equilibrarse. Observando la gráfica 4.2(a) que agrupa por mes, apreciamos que hay una diferencia bastante notoria en los periodos de mayo, junio y julio y de noviembre, diciembre y enero. Por último en la gráfica 4.2(b) donde se agrupa por día de la semana (de 0 a 6 empezando por el lunes) se aprecia



(a) Sentido de los desvíos por mes



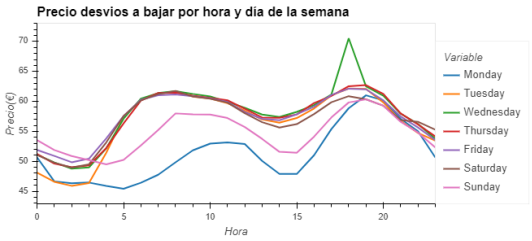
(b) Sentido de los desvíos por día de la semana



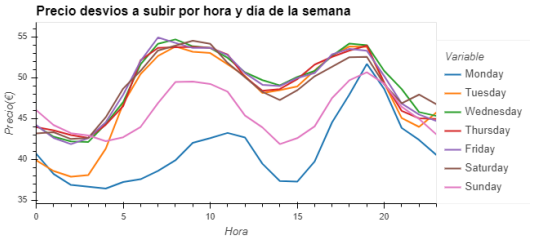
(c) Sentido de los desvíos por hora

Figura 4.2: Distribución del sentido de los desvíos (1 para desvíos a subir y -1 para desvíos a bajar) agrupados por diferentes resoluciones temporales. Fuente: Elaboración propia

una diferencia mucho menor que en otras agrupaciones.



(a) Precio de los desvíos a bajar agrupados por hora y día de la semana

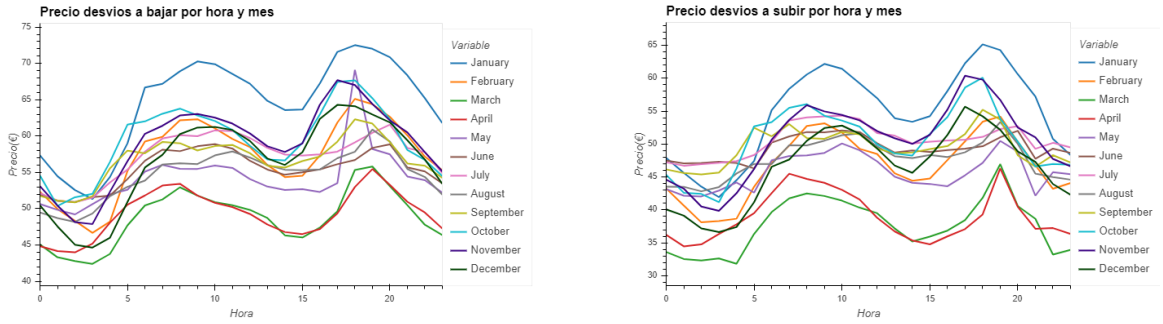


(b) Precio de los desvíos a subir agrupados por hora y día de la semana

Figura 4.3: Precio de los desvíos a subir y a bajar agrupados por hora y día de la semana. Fuente: Elaboración propia a partir de datos de ESIOS

En lo que se refiere al precio de los desvíos, observamos también un gran variabilidad a lo largo del tiempo, como se aprecia en las figuras 4.3 y 4.4. En 4.3 agrupamos el precio de los desvíos por hora y día de la semana, y vemos que la trayectoria temporal es parecida de martes a sábado, siendo los lunes y los domingos días con un comportamiento totalmente distinto. En 4.4 agrupamos por hora y mes, pudiendo ver que enero, marzo y abril siguen un comportamiento notablemente diferente al resto.

Todo lo observado en estas gráficas nos indican la gran influencia que tiene el momento temporal



(a) Precio de los desvíos a bajar agrupados por hora y mes

(b) Precio de los desvíos a subir agrupados por hora y mes

Figura 4.4: Precio de los desvíos a subir y a bajar agrupados por hora y mes. Fuente: Elaboración propia a partir de datos de ESIOS

tanto en el sentido de los desvíos como en el precio. Por esta razón será de gran utilidad incorporar a los patrones de entrenamiento variables temporales.

4.3. Variables utilizadas

En total se han recolectado 49 variables del ESIOS, las cuales abarcan todos ámbitos del sistema eléctrico español, por lo que para listarlas y entenderlas debemos de clasificarlas debidamente. Se agrupan en dos tipos:

- **Variables de "tiempo real"**. Son las que se miden y se proporcionan para la última hora acontecida. Se han escogido todas las variables de generación de las distintas energías renovables y no renovables, la demanda real, el precio de los desvíos o su sentido.
- **Variables "a futuro"**. Son variables que se publican a partir de predicciones generadas por REE. Puesto que nos proporcionan información a futuro, nos resultan de especial utilidad para llevar a cabo nuestras predicciones. Abarcan desde predicciones del estado de los intercambios con otros países, predicción del precio de la energía, predicción de la demanda o incluso la previsión que se da para la generación de energía eólica y fotovoltaica.

Por otro lado, tal y como se ha mencionado en el apartado anterior, se ha hecho uso de datos de los días festivos de España para conocer el porcentaje de población que se encuentra en festivo para un momento determinado del año.

Además, se han aplicado técnicas de *feature engineering* [19] añadiendo variables generadas a partir de otras recolectadas previamente. Puesto que hay una clara estacionalidad en la demanda, en los precios y en los desvíos, se han añadido como variables el valor que estos mismos tenían el día anterior a la misma hora y el valor que tenían la semana anterior el mismo día a la misma hora.

Por último, debido a esta temporalidad que se ha mencionado, se ha creído que sería relevante añadir variables temporales que indiquen el mes del año, la semana del año, el día de la semana, el

día del mes y el día del año. La problemática que presentan este tipo de variables cíclicas es que el último valor del ciclo está a distancia máxima del primero, por lo que para un modelo de predicción se tomarían como valores opuestos. Por este motivo, todas estas variables cíclicas mencionadas son calculadas de la forma:

$$variableSeno = sen((2 * \pi * variable)/nPosiblesValores)$$

$$variableCoseno = cos((2 * \pi * variable)/nPosiblesValores)$$

Por ejemplo, para el caso de los posibles valores de un mes, aplicando las fórmulas mostradas nos quedarían valores como los mostrados en la figura 4.5, en donde los valores más próximos al 12 son el 11 y el 1.

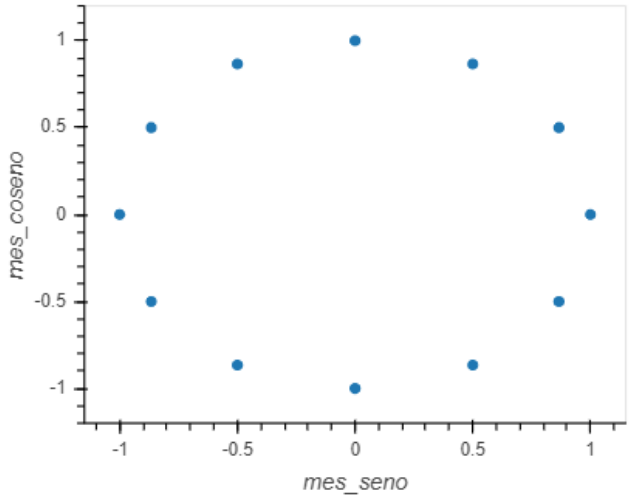


Figura 4.5: Ejemplo de variable temporal añadida al estudio. Fuente: Elaboración propia

4.4. Herramientas empleadas en la implementación

Todo el desarrollo de este trabajo ha sido implementado en *Python*. Actualmente es el lenguaje más extendido dentro del campo de la Ciencia de Datos gracias a las facilidades que ofrece en la codificación al programador y a la gran variedad de librerías diseñadas con este propósito:

- En el **módulo de obtención de datos** se ha empleado la librería *requests*, que nos permite realizar peticiones a las APIs del ESIOS y de Nager.Date.
- Para la parte de **procesado y transformación de los datos** se han empleado librerías como *numpy* [22] y *pandas* [23]. La primera nos permite hacer transformaciones numéricas sobre conjuntos de datos. La segunda está construida sobre *numpy* y nos ayuda a tratar datos tabulares de forma sencilla.
- En el **módulo de predicción** y de test se ha utilizado la librería *scikit-learn* [24]. Es la más utilizada en el campo

del *Machine Learning*, ya que ofrece la implementación de una gran variedad de modelos, transformadores y métricas de evaluación de resultados y una gran comunidad muy activa que constantemente trabaja en actualizarla y mejorarla.

- Para la parte de **visualización gráfica de datos**, se ha empleado la librería *hvPlot* [25], del ecosistema *Holoviz*. Nos permite escoger entre una amplia variedad de estilos de gráficas para representar los datos que se tengan cargados en un *dataframe* de *pandas*.

4.5. Construcción del *dataset*

Una vez se han recolectado y construido las variables mencionadas previamente, se ha procedido a generar el *dataset* que nuestros modelos utilizarán durante el entrenamiento. Para ello, se han desarrollado dos versiones.

En la **versión *multi-output*** se han dispuesto los datos del *dataset* de forma que cada fila corresponde a un día D. Esta fila estará compuesta por todas las variables "de tiempo real" recogidas durante las 24 horas del día D-1, las variables "a futuro" para las 24 horas del día D y nuestro target para las 24 horas del día D. El esquema queda reflejado en la figura 4.6.

Día	Var1_1	Var1_2	...	Var1_24	Var2_1	Var2_2	...	Var2_24	...	Target1	Target2	...	Target24
D	x1_1	x1_2	...	x1_24	x2_1	x2_2	...	x2_24	...	y1	y2	...	y24

Figura 4.6: Esquema de un dataset *multi-output*. Fuente: Elaboración propia

Tendremos tantas filas como días vayamos a usar para entrenar el modelo. En este caso, se han seleccionado datos de dos años completos (2017 y 2018), por lo que nos quedarán alrededor de 730 patrones.

La **versión *single-output*** es más simple, pues en ella tenemos una fila por hora en vez de por día. Para todas las filas correspondientes al día D tendremos los mismos valores para cada variable, ya que, al igual que en el *multi-output* tenemos una columna por cada variable en cada una de las horas. Sin embargo, al añadir las variables temporales descritas en la sección 4.3 *Variables utilizadas*, hacen que cada fila sea distinta según el momento temporal en que se encuentre la hora a predecir. Podemos ver un esquema de la forma de este *dataset* en la figura 4.6.

Datetime	Var1_1	Var1_2	...	Var1_24	Var2_1	Var2_2	...	Var2_24	...	Target
H1	x1_1	x1_2	...	x1_24	x2_1	x2_2	...	x2_24	...	y1
H2	x1_1	x1_2	...	x1_24	x2_1	x2_2	...	x2_24	...	y2
..	..	..	...	..	...	...	...	...	...	...

Figura 4.7: Esquema de un dataset *single-output*. Fuente: Elaboración propia

Por lo tanto, el *dataset* tendrá alrededor de 17.500 patrones, ya que son 730 días * 24 horas.

Datasets de las mismas características serán utilizados para la parte de test de nuestros modelos, con la salvedad de que se les retirará las columnas de target, ya que es lo que queremos predecir. Para esta tarea de test se usarán datos del año 2019 completo, por lo que dispondremos de unos 365 patrones para validar los modelos *multi-output* y 8.700 para los *single-output*.

La ventaja de disponer los datos en modo *multi-output* viene de la mano de que podemos entrenar un modelo distinto por horizonte de predicción. Debido a la estacionalidad observada previamente donde la hora del día influye en los desvíos severamente, tiene sentido tratar cada uno de estos horizontes de manera independiente. Por otro lado, la ventaja del modo *single-output* está en que disponemos de mucho más histórico con el que entrenar los modelos, pasando de tener un patrón por día a un patrón por hora.

4.6. *Pipeline* de predicción

Para llevar a cabo las predicciones deseadas, necesitamos diseñar un pipeline de procesos que nos ayuden a preprocesar los datos, separarlos en diferentes conjuntos de entrenamiento y validación y buscar los hiperparámetros de nuestros modelos que nos ayuden a obtener el mejor rendimiento. Se ha utilizado el *pipeline* que ofrece *scikit-learn* que nos ofrece la ventaja de que, una vez lo tenemos construido, podemos aplicarlo tanto en la fase de entrenamiento como en la de test de forma muy sencilla y flexible.

En lo que se refiere al preproceso de datos, se ha aplicado una **normalización** de los mismos mediante un *Standard Scaler* que consiste en que a cada patrón x de nuestro dataset se le aplica la siguiente fórmula: $z = (x - u)/s$, donde u es la media de todo el conjunto de datos para esa variable, y s es la desviación estándar. Aunque este *scaler*, al emplear la media para su cómputo, no es especialmente bueno frente a una gran cantidad de *outliers*, no será considerado un problema al no habernos encontrado con esta situación mencionada. Además, es necesario efectuar este paso ya que, de no hacerlo, los modelos podrían tomar como representativos datos con valores muy altos o muy bajos, teniendo un impacto negativo en la calidad de nuestra predicción

En la hiperparametrización de nuestros modelos, se ha empleado el algoritmo ***Grid Search*** o *búsqueda en rejilla*. Se basa en que definamos una serie de valores que pueden tomar cada uno de los hiperparámetros, para los que automáticamente se va realizando las combinaciones entre todos ellos en busca del que mejor resultado nos proporcione. El motivo de establecer esta búsqueda la encontramos en que habitualmente los parámetros por defecto de los modelos no suelen dar resultados óptimos, por lo que es conveniente los que mejor se adecúan a nuestro problema. En la configuración del *Grid Search* se han tenido una serie de consideraciones:

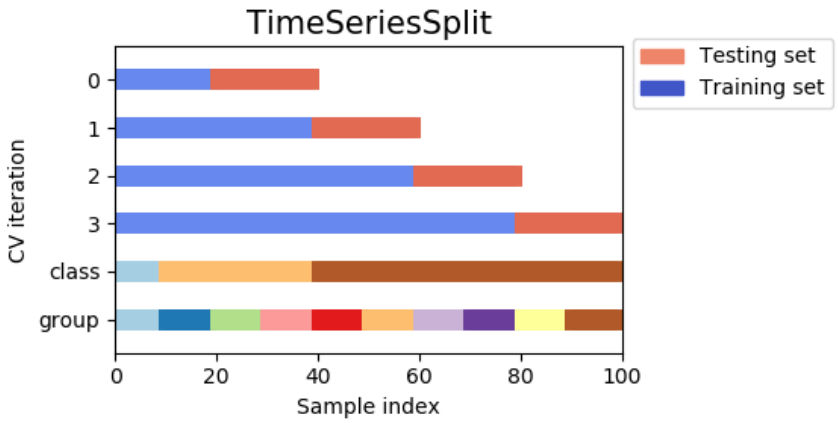


Figura 4.8: Esquema de funcionamiento de *Time Series Split*. Fuente: https://scikit-learn.org/

- Para la **validación** durante la búsqueda en rejilla, se ha utilizado una validación mediante *Time Series Split*, una variación del *k-fold*. Como podemos ver en la figura 4.8, su funcionamiento se basa en que se fraccionan los datos en una parte de entrenamiento y otra de validación, con la característica especial de que el conjunto de validación nunca pertenece a una etapa previa a la de entrenamiento, como sí que ocurriría utilizando *k-fold*. Esto nos permite realizar un entrenamiento realista, pues en ningún momento aplicamos el planteamiento engañoso de que nuestro modelo se entrene con el futuro para predecir el pasado.
- Se ha mencionado que el *Grid Search* busca los mejores parámetros para nuestros modelos, pero necesitamos definir para ello lo que se debe entender por "mejor". En el caso del problema de clasificación, haremos que nuestra búsqueda en rejilla encuentre los parámetros que mejor *F1 score* nos proporcione, ya que esta métrica es, como se ha explicado previamente, la que mejor equilibrio nos da entre *precission* y *recall*. En cuanto al problema de regresión se buscan los que mejor *MAE* nos ofrezcan, debido a que buscamos aquellas predicciones que menos se desvíen del valor correcto. Para más información sobre estas métricas, consultar la sección 3.3.
- La elección de los parámetros ha resultado complicada debido a la gran cantidad de algoritmos probados. Para ayudarnos a realizar correctamente esta tarea, se ha empleado la librería *HiPlot* [26]. Nos permite visualizar de forma interactiva el resultado de nuestras hiperparametrizaciones, de modo que podemos observar si alguno de los mejores parámetros obtenidos se encuentra en el máximo o en el mínimo de los valores entre los que se han probado en la rejilla. En caso de ser así, deberemos volver a realizar la metamodelización ampliando este rango de búsqueda para asegurarnos de tener un parámetro que saque al modelo su máximo potencial. Podemos ver un ejemplo de este tipo de gráficas que se ha utilizado en la figura 4.9.

4.7. Módulo de *test*

Nuestro objetivo final en este trabajo es el de encontrar el modelo que nos ofrezca las mejores predicciones. Para ello, se ha implementado un módulo de testing que nos permite evaluar el rendimiento de los modelos utilizados de una manera bastante sencilla y flexible.

Gracias a este módulo podemos seleccionar todos los modelos entrenados que deseemos, aplicarles un dataset de test (distinto al que se ha usado para entrenar) y una resolución temporal que deseemos (por hora, mes, semana del año, día de la semana, día del año o fecha). De este modo, ob-

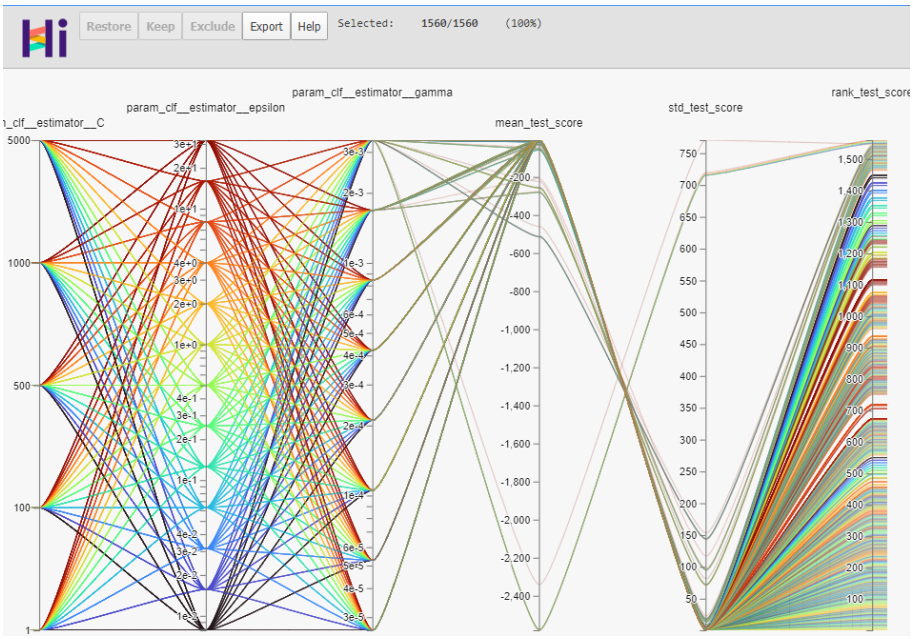


Figura 4.9: Ejemplo de gráfica de test de rendimiento de una metamodelización mediante *GridSearch*. Fuente: Elaboración propia.

tendremos un dataset de los resultados de aplicar las métricas explicadas en 3.3 a todos los modelos, agrupándolos por la resolución temporal elegida.

Este dataset podrá ser utilizado para realizar todas las gráficas de resultados que necesitemos y poder escoger así fácilmente nuestra mejor solución al problema planteado.

5

Pruebas y resultados

5.1. *Benchmark* de modelos

El propósito de esta sección es el de mostrar el benchmark de modelos realizado con el fin de realizar las pruebas necesarias para encontrar el mejor resultado en nuestras predicciones. Se han desarrollado una serie de módulos en Python que ejecutan secuencialmente la metamodelización y modelización de los diferentes modelos empleados utilizando los datos de 2017 y 2018. Debido a la gran cantidad de tiempo computacional que requiere esta tarea, conseguimos mediante estos scripts dejar todo programado para que se ejecute todo durante días.

Cabe aclarar que predicción del precio de los desvíos se ha efectuado únicamente para los "desvíos a subir", ya que, como hemos visto, tienen un comportamiento muy similar a los "desvíos a bajar".

5.1.1. *Benchmark* de modelos de clasificación

En esta sección el objetivo es el de predecir el sentido de los desvíos del mercado (ascendente o descendente), por lo que es sin duda un problema de clasificación. Para esta tarea se han desarrollado dos scripts distintos dependiendo del diseño de dataset que se vaya a utilizar: *multi-output* o *single-output*. Para más información sobre cómo son estos *datasets*, consultar la sección *Construcción del dataset* 4.5.

- Por el lado del ***single-output*** el módulo contiene la metamodelización del *Logistic Regression*, *SVM*, *Random Forest* y *Multi Layer Perceptron*.
- En cuanto el caso del ***multi-output*** se han probado también *Logistic Regression*, *SVM*, *Random Forest* y *Multi Layer Perceptron*, con la consideración de que estos dos últimos poseen por sí mismos las características necesarias para proporcionar un resultado *multi-output* sin necesidad de utilizarlos sobre un *multi-output classifier* donde se realiza un modelo distinto por horizonte de predicción. Por esta razón, probaremos el desempeño de estos modelos en ambas versiones.

5.1.2. *Benchmark* de modelos de regresión

El objetivo de esta sección es el de predecir el precio de los desvíos, por lo que claramente es un problema de regresión. Al igual que ocurre en el *benchmark* de clasificación, se realizarán dos scripts distintos dependiendo de si estamos utilizando el *dataset multi-output* o *single-output*.

- En lo que se refiere al ***single-output***, se aplican *SVM*, *Random Forest* y *Multi Layer Perceptron* en su versión de regresión.
- Para el ***multi-output***, se han probado los mismos modelos a los que se les añade, al igual que en clasificación las versiones de *Random Forest* y *Multi Layer Perceptron* sin ser utilizadas sobre un *multi-output regressor* con el objetivo de probar el rendimiento de ambas versiones.

5.2. Resultados de metamodelización

Como se ha explicado en la sección 4.6, se ha llevado a cabo un proceso de metamodelización mediante el algoritmo *Grid Search* de todos los modelos empleados en el *benchmark* en busca de los parámetros que mejor resultados nos ofrezcan. En esta sección se recopilan los mejores parámetros obtenidos para cada uno de nuestros modelos:

- Predicción del sentido de los desvíos con *multi-output*:
  - ***Logistic Regression*** con *MultioutputClassifier*: C=0.1 y tolerance=0.1.
  - ***SVM*** con *MultioutputClassifier*: C=4000, gamma=2.75e-06.
  - ***Random Forest*** con *MultioutputClassifier*: n-estimators=3000.
  - ***Random Forest***: n-estimators=4000.
  - ***MLP*** con *MultioutputClassifier*: activation='tanh', alpha=1.
  - ***MLP***: activation='relu', alpha=0.1.
- Predicción del sentido de los desvíos con *single-output*:
  - ***Logistic Regression*** con *MultioutputClassifier*: C=0.1 y tolerance=1.
  - ***SVM*** con *MultioutputClassifier*: C=1, gamma=1.34e-05.
  - ***Random Forest***: n-estimators=2000.
  - ***MLP***: activation='tanh', alpha=0.1.
- Predicción del precio de los desvíos con *multi-output*:
  - ***SVM*** con *MultioutputRegressor*: C=500, gamma=2.106e-03, epsilon=1.997.
  - ***Random Forest* con *MultioutputRegressor***: n-estimators=2000.
  - ***Random Forest***: n-estimators=3000.
  - ***MLP* con *MultioutputRegressor***: activation='relu', alpha=0.0001.
  - ***MLP***: activation='relu', alpha=0.0001.
- Predicción del precio de los desvíos con *single-output*:
  - ***SVM***: C=2000, gamma=0.003, epsilon=0.499.
  - ***Random Forest***: n-estimators=1000.

- ***MLP***: activation='relu', alpha=0.001.

5.3. Resultados de predicción

Una vez entrenados los modelos detallados en la sección sobre el *Benchmark de modelos* 5.1, llevaremos a cabo la fase de *test*, es decir, la fase mediante la cual medimos la capacidad de los modelos de predecir correctamente el resultado deseado. Para ello, se han proporcionado a todos los modelos entrenados los datos referentes a todo el año 2019 con el objetivo de medir cuál hubiese sido su desempeño durante dicho año.

En esta sección, se recogerán y visualizarán los resultados de los modelos de clasificación aplicados a la predicción del precio de los desvíos y los de los modelos de regresión aplicados a la predicción del precio.

Debido a que, como se ha mencionado previamente, no se han podido encontrar otros trabajos en los que se trate este mismo tema con el que poder comparar los resultados obtenidos, se han establecido una serie de métricas que nos pueden servir como referencia. Estas métricas están basadas en el concepto de "persistencia", que es un término ampliamente empleado en el mundo del *Machine Learning* para denominar a aquella predicción efectuada simplemente replicando un resultado real del pasado. Utilizaremos:

- **Día-hora anterior** consiste en predecir para la hora H del día D+1, lo acontecido a la hora H del día D.
- **Semana-día-hora anterior** consiste en predecir para la hora H del día D de la semana S+1, lo acontecido a la hora H del día D de la semana S.
- **Persistencia siempre -1**, empleado solo en clasificación, consiste en predecir siempre nuestra clase mayoritaria.

5.3.1. Resultados de predicción del sentido de los desvíos

En la tabla 5.1 podemos observar el resultado de aplicar las métricas de *precision*, *recall* y *F1* a las predicciones efectuadas durante la fase de *test* utilizando **modelos *multi-output***. Aunque podemos observar que no hay una diferencia muy significativa entre los distintos modelos, nos debemos quedar con la *SVM*, ya que es el modelo que mejor *F1* nos aporta. Esta medida la podemos considerar la más relevante, ya que, como ya se explicó en 3.3, nos aporta el mejor equilibrio entre *precision* y *recall* . Además, es el que mayor *precision* muestra, que nos mide la capacidad de no etiquetar como A un elemento de clase B.

Si comparamos los resultados obtenidos con las medidas de persistencia que hemos establecido como base, vemos que nuestras predicciones presentan un mejor desempeño por una diferencia bastante significativa.

En la figura 5.1 podemos ver la evolución del *F1 score* a lo largo de los distintos meses de 2019

para los que se realiza la predicción, apreciando que la calidad de la predicción para todos los modelos es variable a lo largo de los meses. A pesar de que hemos visto que la *SVM* es el modelo que mejor resultados nos ofrece de media, vemos que hay muchos meses en los que el *Random Forest* le supera.

Model	Precision	Recall	F1
LR con MultioutputClassifier	0.728	0.631	0.645
SVM con MultioutputClassifier	0.731	0.633	0.648
RF con MultioutputClassifier	0.712	0.627	0.635
RF	0.713	0.643	0.645
MLP con MultioutputClassifier	0.704	0.604	0.627
MLP	0.707	0.595	0.616
Persistencia dia-hora anterior	0.604	0.605	0.606
Persistencia semana-día-hora anterior	0.546	0.546	0.548
Persistencia siempre -1	0.287	0.364	0.5

Tabla 5.1: Tabla resultados de predicción del sentido de los desvíos aplicando modelos de clasificación para *multi-output*

En cuanto a los resultados obtenidos en *test* para el caso de **modelos *single-output***, obtenemos los valores que quedan reflejados en la tabla 5.2. Podemos ver que una vez más la diferencia es mínima, pero la SVM vuelve a ser el modelo que mejor resultados nos ofrece.

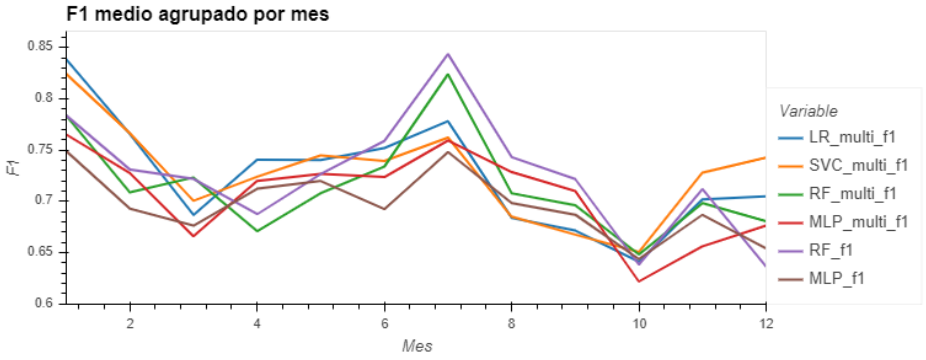


Figura 5.1: Gráfica del *F1 score* agrupado por mes en *multi-output*. Fuente: Elaboración propia

Sin embargo, si nos fijamos en la figura 5.2, vemos también, tal y como ocurría en la versión de *multi-output*, hay meses en los que otros modelos demuestran ser mejores que la *SVM*.

Teniendo en cuenta todo lo mencionado previamente, podemos llegar a concluir que la versión *multi-output* nos ofrece una mejor solución al problema planteado de predicción del sentido de los desvíos que la versión *single-output*. Además, hemos visto que la *SVM* es el modelo más regular en sus predicciones a lo largo del año.

En el caso de que fuesemos capaces de desarrollar un sistema encargargado de elegir el modelo

Model	Precision	Recall	F1
LR	0.637	0.591	0.604
SVM	0.647	0.637	0.652
RF	0.641	0.636	0.645
MLP	0.596	0.533	0.544
Persistencia dia-hora anterior	0.604	0.605	0.606
Persistencia semana-día-hora anterior	0.546	0.546	0.548
Persistencia siempre -1	0.287	0.364	0.5

Tabla 5.2: Tabla resultados de predicción del sentido de los desvíos aplicando modelos de clasificación para *single-output*

a utilizar durante las distintas fases del año, podríamos llegar a tener unas predicciones aún más acertadas, pues nos beneficiaríamos de los puntos fuertes que se han visto que tiene cada modelo en cada una de dichas fases.

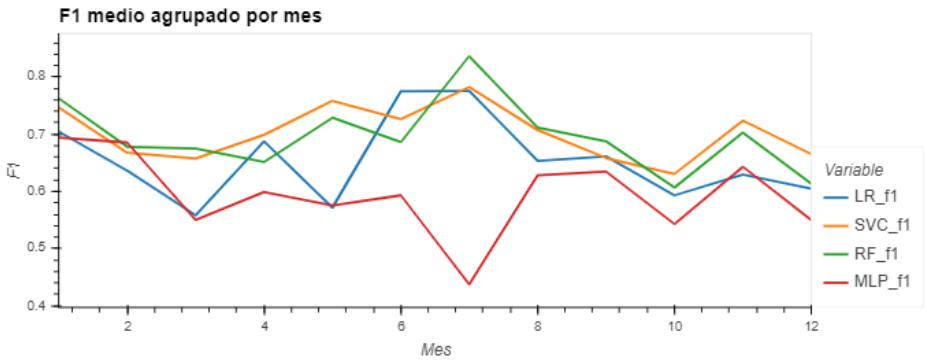


Figura 5.2: Gráfica del *F1 score* agrupado por mes en *single-output*. Fuente: Elaboración propia

5.3.2. Resultados de predicción del precio de los desvíos

Para el problema de la predicción del precio de los desvíos, se aplican otro tipo de métricas, ya que es un problema de regresión y no de clasificación. Podemos observar en la tabla 5.3 el resultado de la fase de *test* utilizando la versión de **modelos *multi-output***. Observamos que el mejor resultado nos lo ha ofrecido la *SVM*, que tiene 5.32 de *MAE*, es decir, que de media tiene un error de 5.32€ respecto al precio finalmente establecido.

Comparando los resultados logrados con las medidas de persistencia, observamos que nuestras predicciones presentan también un mejor desempeño tanto en el *MAE* como en el *MSE*.

Observando la figura 5.3 del *MAE* por mes, apreciamos que, a pesar de que la *SVM* es el que da mejores resultados en el cómputo global del año, hay meses donde el *MLP* que ofrece un error menor.

Model	MAE	MSE
SVM con MultioutputRegressor	5.321	64.553
RF con MultioutputRegressor	6.202	84.008
RF	6.344	95.110
MLP con MultioutputRegressor	6.725	98.065
MLP	6.894	98.645
Persistencia con día-hora anterior	8.558	155.429
Persistencia con semana-día-hora anterior	10.792	230.917

Tabla 5.3: Tabla resultados de predicción del precio de los desvíos aplicando modelos de regresión para *multi-output*

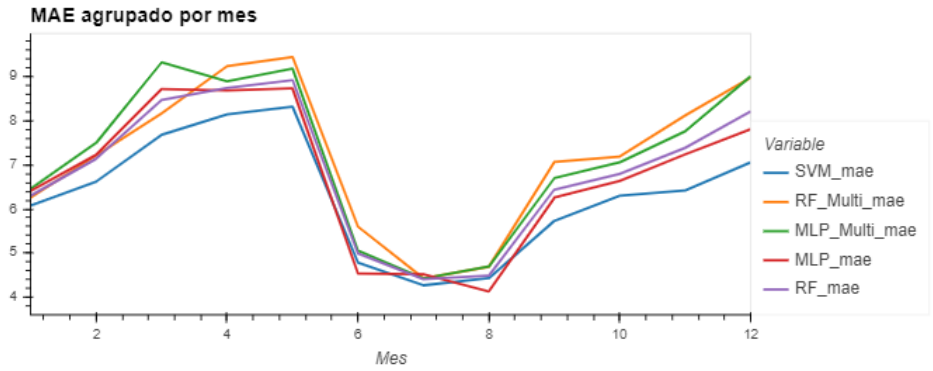


Figura 5.3: Gráfica del *MAE* agrupado por mes en *multi-output*. Fuente: Elaboración propia

Model	MAE	MSE
SVM	5.360	60.978
RF	6.251	80.176
MLP	7.131	91.371
Persistencia con día-hora anterior	8.558	155.429
Persistencia con semana-día-hora anterior	10.792	230.917

Tabla 5.4: Tabla resultados de predicción del precio de los desvíos aplicando modelos de regresión para *single-output*

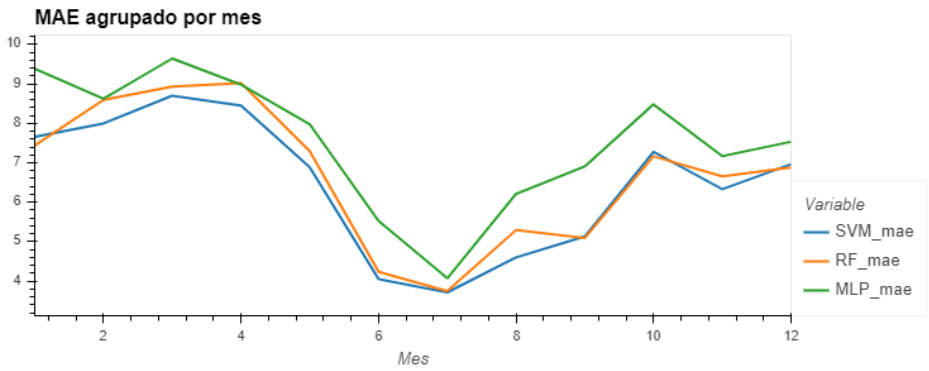


Figura 5.4: Gráfica del *MAE* agrupado por mes en *single-output*. Fuente: Elaboración propia

En lo que se refiere a la versión de **modelos *single-output***, vemos en la tabla 5.4 unos resultados donde también la SVM vuelve a ofrecer la *SVM*, con un *MAE* parecido al visto anteriormente.

En la gráfica 5.4, vemos que la *SVM* demuestra ser el modelo que presenta menor error casi para la totalidad de los meses predichos.

En resumen, en el caso de la predicción del precio, deberemos utilizar también la *SVM*. Aunque las versiones *multi-output* y *single-output* ofrecen un *MAE* muy similar, deberemos elegir la segunda, ya que ofrece un *MSE* sustancialmente mejor, por lo que entendemos que es más tolerante a los fallos. Por otro lado, tal y como hemos visto que sucede también en la predicción del sentido de los desvíos, hay modelos que presentan mejores resultados para ciertos meses del año, por lo que puede llegar a ser recomendable su alternancia en el tiempo para poder obtener así los mejores resultados.

6

Conclusiones y trabajo futuro

6.1. Conclusiones

Con este trabajo, se han podido conocer y explorar en profundidad las distintas fases que componen un proyecto del ámbito de la Ciencia de Datos, según la metodología CRISP-DM.

En primer lugar, ha sido necesario conocer bien el campo en el que se encuadra este trabajo: el mercado eléctrico español, y más concretamente el mercado de gestión de *desvíos*. Esto nos ha permitido entender el comportamiento de este fenómeno y tener una percepción clara sobre qué variables pueden explicarlo.

Realizar un módulo de obtención de datos de forma dinámica, nos ha permitido disponer durante el desarrollo del proyecto de una herramienta flexible mediante la cual podíamos incorporar de forma sencilla cualquier nueva variable que creyéramos que podría ser útil. Además, esto se ha complementado muy bien con los *scripts* que componen el estudio estadístico de los datos, pues podíamos ver y entender rápidamente todas las variables de las que se dispone.

Se han implementado un *benchmark* de modelos y técnicas de *Machine Learning* con el fin de obtener el mejor predictor, tanto del sentido de los desvíos, como de sus precios. La combinación de ambos problemas, tan distintos entre sí al ser uno de clasificación y otro de regresión, nos ha permitido realizar un recorrido bastante amplio sobre el campo del *Machine Learning*. Además, el poder haber habido experimentar acerca de diferentes métodos de presentar los datos a los modelos (*single-output* y *multi-output*), ha provocado que obtengamos unas predicciones con un mejor acierto.

Asimismo, el haber implementado un módulo de test flexible, también nos ha supuesto una gran ayuda a la hora de probar el desempeño de nuestros modelos. Debido a que teníamos modelos de clasificación y de regresión, que a su vez usaban *single-output* y *multi-output*, era importante disponer de un módulo que de forma dinámica se adaptase a cada una de estas casuísticas.

Gracias a la extensa comparativa realizada en el *benchmark* de modelos, hemos podido establecer que el modelo que mejor resuelve nuestro problema de clasificación del sentido de los desvíos es la *SVM* en su versión *multi-output* con un 0.65 de *F1 score* y 0.73 de *precision score*. En el caso de la

predicción del precio de los desvíos, también ha resultado ser la *SVM* en su versión *multi-output* el que mejor rendimiento nos ofrece, con un *MAE* de 5.32€. Estos resultados obtenidos pueden ser tomados como *baseline* para posteriores investigaciones y desarrollos sobre esta cuestión.

Los modelos implementados pueden llegar a ser ser de gran utilidad para los agentes del mercado eléctrico, especialmente si a los modelos se les incorpora datos acerca de la estrategia que van a seguir dichos agentes. Todas sus decisiones tienen un gran impacto en el mercado, especialmente en los desvíos. Por lo tanto, conocer de antemano la manera de operar de los agentes puede suponer una gran mejora en los resultados de nuestro problema.

6.2. Trabajo futuro

El precio de los desvíos obedece en gran parte a las variables que se han tenido en cuenta a lo largo de este estudio. Sin embargo, con el objetivo de afinar estas predicciones, podríamos estudiar la incorporación de nuevas variables a nuestros modelos:

- La predicción de la producción de la energía eólica y fotovoltaica es un dato que aporta ya el ESIOS, pero no ha podido tenerse en cuenta debido a que solo se lleva haciendo desde 2019, por lo que no tenemos los suficientes datos para entrenar los modelos.
- Datos acerca de otros países interconectados con España, puesto que pueden llegar a tener gran relación en la necesidad del sistema convocar el mercado de desvíos.
- Datos de predicción eólica generada por otros agentes predictores.
- Datos sobre los planes estratégicos llevados a cabo por las empresas generadoras de energía. Sin embargo, este tipo de datos acostumbran a ser privados.
- La realimentación con predicciones efectuadas por el mismo modelo desde horizontes de predicción anteriores.

También se puede trabajar en mejorar las variables temporales aportadas a este estudio de forma que sean más precisas, ya que actualmente solo se tienen en cuenta los festivos nacionales, autonómicos y regionales, pero no a nivel de municipio. Tampoco se considera la importancia de esas fiestas o el carácter que tienen.

Por otro lado, se ha visto que la mejor predicción la obtendríamos si fuesemos capaces de alternar inteligentemente entre los distintos meses del año. Por esta razón, una linea de mejora podría venir de mano de la implementación de un módulo que fuese capaz de automatizar este proceso.

Por último, la ampliación del trabajo realizado puede estar en el estudio de nuevos modelos y técnicas de predicción. Especialmente en la parte de predicción de los precios, que es un problema de regresión al uso, para el que pueden encontrarse una gran cantidad de soluciones publicadas.

Bibliografía

[1] "REE (Red Eléctrica de España)." https://www.ree.es/es. [Online; accessed 10-March-2020].

[2] "Descripción a cerca de la metodología CRISP-DM." https://www.sngular.com/es/data-science-crisp-dm-metodologia/. [Online; accessed 26-April-2020].

[3] "Energía y Sociedad." http://www.energiaysociedad.es/quienes-somos/. [Online; accessed 10-March-2020].

[4] "ESIOS (Sistema de Información del Operador del Sistema)." https://www.esios.ree.es/es. [Online; accessed 10-March-2020].

[5] "BOE: Ley 54/1997." https://www.boe.es/buscar/doc.php?id=BOE-A-1997-25340. [Online; accessed 15-April-2020].

[6] "MIBEL (Mercado Ibérico de la Electricidad)." https://www.mibel.com/. [Online; accessed 10-March-2020].

[7] "OMIE (Operador del Mercado Ibérico de Energía - Polo Español)." https://www.omie.es/. [Online; accessed 10-March-2020].

[8] "Orden de agregación de precios por tecnología." https://aleasoft.com/es/funcionamiento-mercado-electrico-iberico-mibel/. [Online; accessed 23-April-2020].

[9] "Compromiso de REE de equilibrar generación y consumo." https://www.ree.es/es/actividades/operacion-del-sistema-electrico. [Online; accessed 23-April-2020].

[10] "Why Data Science Succeeds or Fails." https://towardsdatascience.com/why-data-science-succeeds-or-fails-c24edd2d2f9. [Online; accessed 2-May-2020].

[11] A. Sanchez Salas and M. Kessler, "Análisis de la predictibilidad de los desvíos eléctricos en el sistema eléctrico español usando redes neuronales," *Universidad Politécnica de Cartagena*, 2015. (Descargar).

[12] F. van Daalen, E. Smirnov, N. Davarzani, R. Peeters, J. Karel, and H. Brunner-La Rocca, "An ensemble approach to time dependent classification," *Maastricht UMC*, 2018. (Descargar).

[13] J. Read, L. Martino, P. Olmos, and D. Luengo, "Scalable multi-output label prediction: From classifier chains to classifier trellises," *AaltoUniversity, University of Helsinki, Universidad Carlos III de Madrid, Universidad Politécnica de Madrid*, 2015. (Descargar).

[14] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[15] A. Romero, J. R. Dorronsoro, and J. Díaz, "Day-ahead price forecasting for the spanish electricity market," *International Journal of Interactive Multimedia and Artificial Intelligence 5(4)*, 2019. (Descargar).

[16] A. Conejo, M. Plazas, and A. Molina, “Day-ahead electricity price forecasting using the wavelet transform and arima models,” *IEEE*, 2005. (Descargar).

[17] M. Opper and O. Winther, “Gaussian processes and svm: Mean field results and leave-one-out,” 1999. (Descargar).

[18] H. Bourlard and Y. Kamp, “Auto-association by multilayer perceptrons and singular value decomposition.,” 1988. (Descargar).

[19] P. Bruce and A. Bruce, *Practical Statistics for Data Scientists*. O’REILLY, 2017.

[20] “Nager.Date API.” `https://date.nager.at/Api`. [Online; accessed 1-March-2020].

[21] “INE (Instituto Nacional de Estadística).” `https://www.ine.es/`. [Online; accessed 1-March-2020].

[22] “Numpy (Librería de *Python*).” `https://numpy.org/`. [Online; accessed 9-April-2020].

[23] “Pandas (Librería de *Python*).” `https://pandas.pydata.org/docs/`. [Online; accessed 9-April-2020].

[24] “Scikit-Learn (Librería de *Python*).” `https://scikit-learn.org/stable/`. [Online; accessed 9-April-2020].

[25] “hvPlot (Librería de *Python* del ecosistema *Holoviz*).” `https://hvplot.holoviz.org/`. [Online; accessed 9-April-2020].

[26] “HiPlot (Librería de *Python*).” `https://pypi.org/project/hiplot/`. [Online; accessed 12-April-2020].


UAM
UNIVERSIDAD AUTONOMA
DE MADRID